**Finding Optimal Stations for NYY Hat Sales**

By: William Paulson

**Abstract**

The goal of this project was to perform exploratory data analysis to search for stations with high traffic before Yankee Home games to help maximize the sales of Yankee hat hawkers. I worked with turnstile data provided by the [MTA (Links to an external site)](http://web.mta.info/developers/turnstile.html) and Yankee schedule information found on [MLB.com (Links to an external site)](https://www.mlb.com/yankees/schedule/2019-06/list).These datasets allowed me to leverage turnstile traffic information and Date/Time information to find the highest traffic subway station during the window of expected to travel to Yankee home games. 161st-Yankee Stadium Station exhibited the most consistent high exit traffic increase during game days compared to regular weekdays. The traffic increase was at its most pronounced on Fridays.

**Design**

This project originates from the Hat Hawkers House (HHH). The turnstile data from the [MTA (Links to an external site)](http://web.mta.info/developers/turnstile.html) allows for quantification of traffic of New York City subway traffic. Finding high traffic stations in specific time windows would enable the Hat Hawkers House to deploy their workforce more efficiently to improve sales and increase company profits.

**Data**

The MTA dataset contained 2,068,262 entries with 10 categorical features. This data encompassed 12 weeks of turnstile data over the course of 4 months. The most important features were station identifiers, cumulative entries, cumulative exits, date, and time information. Grouping by station allowed for the data to be anaylzed at a higher level and provide actionable information.

**Algorithms**

**Exploratory Data Analysis**

1. Read raw data into Python and SQL to build database
2. Query database to find highest exit traffic
3. Analyze the data in Python to understand results
4. Clean the data of outliers, errors, and other issues
5. Aggregate data to make more broad conclusions
6. Visualize data using Matplotlib

I chose to focus the analysis on week-day night games during summer to create the highest density of game days to non-game days while also minimizing the effect of inconsistent traffic patterns due to seasonality, different time frames, and different weekend patterns.

**Tools**

* Pandas and Datetime for data manipulation
* SQLite and SQLAlchemy for querying
* Matplotlib for plotting

**Communication**

In addition to the slides and visuals presented, [MTA turnstile data (Links to an external site)](http://web.mta.info/developers/turnstile.html) will be embedded on my personal website and blog.

Chart, line chart, scatter chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generated